Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

MAS

Modelling, Analysis and Simulation

Modelling, Analysis and Simulation

On parameter estimation and determinability for the model of pattern formation in Drosophila melanogaster

M. Ashyraliyev, J.G. Blom

REPORT MAS-E0802 JANUARY 2008

Centrum voor Wiskunde en Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3703

On parameter estimation and determinability for the model of pattern formation in Drosophila melanogaster

ABSTRACT
Mathematical modelling of real-life processes often requires the estimation of unknown parameters. Once the parameters are found by means of optimization, it is important to asses the quality of the parameter estimates. In this paper we describe how the quality of these estimates can be analyzed and this methodology is applied to study the model for the genetic regulatory network in the Drosophila embryo during the early developmental stages.

2000 Mathematics Subject Classification: 92C15
Keywords and Phrases: parameter estimation; parameter determinability; regulatory genetic networks; Drosophila
Note: This work was carried out under Dutch Bsik/BRICKS project and NWO's `Computational Life Science' program, projectnr. 635.100.010

On parameter estimation and determinability for the model of pattern formation in *Drosophila melanogaster*

M. Ashyraliyev* & J.G. Blom
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

Mathematical modelling of real-life processes often requires the estimation of unknown parameters. Once the parameters are found by means of optimization, it is important to asses the quality of the parameter estimates. In this paper we describe how the quality of these estimates can be analyzed and this methodology is applied to study the model for the genetic regulatory network in the *Drosophila* embryo during the early developmental stages.

2000 Mathematics Subject Classification: primary 92C15;
Keywords and Phrases: parameter estimation; parameter determinability; regulatory genetic networks; Drosophila

1 Introduction

Many real-life processes can be modelled by Ordinary Differential Equations (ODEs) or Partial Differential Equations (PDEs). For instance, in developmental biology, systems of reaction-diffusion equations are used to model spatio-temporal patterns of protein concentrations [1]. A common difficulty is that the model equations usually have a large number of unknown parameters, such as diffusion coefficients, decay and reaction rates, etc. Sometimes missing parameters can be estimated experimentally, but this is rather exceptional. Mostly, it is impossible to find missing parameter values directly. However, usually one can measure other quantities involved in the model. For instance, experimentalists can measure protein or mRNA concentrations. The unknown model parameters can then be found by parameter estimation techniques such that the solution of the mathematical model fits the measured data.

There exists a number of different optimization techniques for parameter estimation. The choice of the technique usually depends on the type of model equations (deterministic or stochastic), as well as on the level of noise in the data. When the model is deterministic and the data is not too noisy, gradient-based methods are efficient optimizers [2]. In this paper we use the Levenberg-Marquardt (LM) method for that purpose. It is a local search approach, meaning that a sufficiently good initial guess for the parameter values is needed. If available, such values can for example be obtained from literature. Otherwise, the LM

* *E-mail address:* M.Ashyraliyev@cwi.nl

method has to be combined with some global search method, such as simulated annealing, a genetic algorithm, an evolution strategy, etc.

Once the parameter estimates have been computed, it is very important to know how reliable they are. For this, confidence regions can be determined. They allow us to assess the quality of the parameter estimates. Ideally, one would wish to determine all parameters accurately enough. In practice, however, this is usually not possible and one has to face an uncertainty in the parameter values. This can be due to insufficient or noisy data or simply because the 'wrong' model is used. In this paper, we do not focus on the latter aspect, assuming that the 'right' model is available.

Cell differentiation and body plan formation of animals occur in embryos at the early developmental stages [3]. The process of cell differentiation is initiated by different morphogen gradients which provide the spatial information by dividing the embryo in different regions. This is followed by the formation of concentration gradients of gene products which are responsible for body plan formation. The process of pattern formation is based on the regulatory interactions among genes and gene products involved in genetic regulatory networks. Mathematical modelling of the correct spatio-temporal pattern formation of gene product concentrations helps to reveal the regulatory interactions among genes as well as to have insight into the dynamics of the underlying processes. In this work, we consider the gap gene system of *Drosophila melanogaster* (fruit-fly). The mathematical model for this system is introduced in [4] and parameter estimation has been used in [5]-[7] by means of global optimization methods. We apply the LM method to estimate the unknown parameters and we study how well these estimates can be determined, based on the available experimental data [8]. Note that the methodology used is generally applicable for a broad range of models, also arising in other fields.

The paper is organized as follows. In Section 2 we describe the theory needed for the parameter estimation problem, with the focus on the gradient-based LM method, and for the statistical analysis which is applied to investigate the quality of the estimates obtained. In Section 3, we study the biological problem concerning the early stage of development of *Drosophila*. The paper is concluded with remarks in Section 4.

2 Theory

We consider a model given by the system of ODEs of the form:

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{f}(t, \mathbf{y}, \theta), & 0 < t \leq T, \\ \mathbf{y}(t, \theta) = \mathbf{y}_0(\theta), & t = 0. \end{cases} \tag{2.1}$$

Here the m-dimensional vector θ contains all unknown parameters, $\mathbf{y}$ is an n-dimensional state vector, and $\mathbf{f}$ is a given vector function, differentiable with respect to t, $\mathbf{y}$ and θ. When components of the initial state vector $\mathbf{y}_0$ are not known, they are considered as unknown parameters, so $\mathbf{y}_0$ may depend on θ. In this work, we assume that (2.1) is the 'right' model for the problem we are interested in. Let us explain what we mean by a 'right' model. Firstly, it implies that (2.1) is a sufficiently accurate mathematical description approximating reality. This means that all relevant knowledge about the processes is incorporated correctly in the vector function $\mathbf{f}$. Thus, the only uncertainty in (2.1) is the vector of unknown parameters θ. Secondly, it means that there exists a 'true' value θ^* for the parameters θ such that

(2.1) represents reality. So, in principle, all unknown parameters can be determined when sufficient and accurate enough data is available.

Remark 2.1 If the model is given by a system of PDEs, then by applying a spatial discretization, it can be reduced to (2.1). However, in such a case one has to be careful with the choice of the grid size of the spatial discretization. On the one hand, the grid should be fine enough, so that the numerical errors introduced by spatial discretization are negligible in comparison with the level of noise in the data. On the other hand, requiring an extremely fine grid would increase the size of the system (2.1). The latter may be crucial in terms of computational complexity.

Let us assume that for (2.1) there are N measurements available. Each measurement, which we denote by $\tilde{y}_i$, is specified by the time t_i when the c_i-th component of the state vector $\mathbf{y}$ is measured. The corresponding model value obtained from (2.1) is denoted by $y_{c_i}(t_i, \theta)$. The above assumptions imply that the difference $|\tilde{y}_i - y_{c_i}(t_i, \theta^*)|$ is solely due to experimental error. We denote the vector of discrepancies between the theoretical values and the measured values by $\mathrm{Y}(\theta)$. Then the least squares estimate $\hat{\theta}$ of the parameters is the value of θ that minimizes the sum of squares

$$S(\theta) = \sum_{i=1}^{N} (y_{c_i}(t_i, \theta) - \tilde{y}_i)^2 = \mathrm{Y}^T(\theta)\mathrm{Y}(\theta), \tag{2.2}$$

see [15, 16]. We note that (2.2) is an appropriate measure under certain assumptions, which we will discuss in Section 2.2. Other measures might be used when these assumptions do not hold.

2.1 Parameter estimation by the Levenberg-Marquardt method

In general, any gradient-based optimization procedure seeks a correction $\delta\theta$ for the parameter vector, such that $S(\theta + \delta\theta) \leq S(\theta)$ holds. The LM method [10] determines the correction as the solution of the equations

$$\left(J^T(\theta)J(\theta) + \lambda I_m\right)\delta\theta = -J^T(\theta)\mathrm{Y}(\theta), \tag{2.3}$$

where $\lambda \geq 0$ is some constant, I_m is the identity matrix of size m and the Jacobian $J(\theta) = \frac{\partial \mathrm{Y}(\theta)}{\partial \theta}$ is the so-called 'sensitivity' matrix of size $N \times m$. The entry $J_{i,j}$ in $J(\theta)$ shows how sensitive the model response is at the i-th data point for a change in the j-th parameter. The LM method can be seen as the combination of two gradient-based approaches: Gauss-Newton and steepest descent. If $\lambda = 0$ in (2.3), it coincides with the Gauss-Newton method. However, when the matrix $J^T(\theta)J(\theta)$ is (almost) singular, to solve (2.3), λ has to be positive and for large λ the LM method approaches the steepest descent method. During the optimization λ is adapted such that the algorithm strives to exploit the fast convergence of the Gauss-Newton method whenever this is possible [10, 11].

In order to solve (2.3), the singular value decomposition (SVD) of the matrix $J(\theta)$ can be used, i.e.

$$J(\theta) = U(\theta)\ \Sigma(\theta)\ V^T(\theta), \tag{2.4}$$

where $U(\theta)$ is an orthogonal matrix of size $N \times m$, such that $U^T(\theta)U(\theta) = I_m$, $V(\theta)$ is an orthogonal matrix of size $m \times m$, such that $V^T(\theta)V(\theta) = V(\theta)V^T(\theta) = I_m$, and $\Sigma(\theta)$ is a

diagonal matrix of size $m \times m$ which contains all singular values σ_i in non-increasing order. Then the correction $\delta\theta$ can be found as

$$\delta\theta = -V(\theta)\ \left(\Sigma^2(\theta) + \lambda I_m\right)^{-1} \Sigma(\theta)\ U^T(\theta)\ \mathrm{Y}(\theta). \tag{2.5}$$

Later, when we study the reliability of the parameters computed, the SVD will play an important role again.

In order to execute an LM optimization step, the vector of discrepancies $\mathrm{Y}(\theta)$, the matrix $J(\theta)$ and its SVD have to be evaluated for each new estimate of θ. For this purpose, one needs to resolve (2.1) for Y and the additional system of variational equations for the entries of J,

$$\begin{cases} \dfrac{\partial}{\partial t}\dfrac{\partial \mathbf{y}}{\partial \theta_i} = \dfrac{\partial \mathbf{f}}{\partial \theta_i} + \dfrac{\partial \mathbf{f}}{\partial \mathbf{y}}\dfrac{\partial \mathbf{y}}{\partial \theta_i}, & 0 < t \le T, \\ \dfrac{\partial \mathbf{y}(t,\theta)}{\partial \theta_i} = \dfrac{\partial \mathbf{y}_0(\theta)}{\partial \theta_i}, & t = 0, \end{cases} \tag{2.6}$$

for $i = 1, 2, \ldots, m$. We note that the costs for performing the SVD and computing the correction (2.5) are negligible in comparison with the computational costs for solving (2.1) and (2.6).

Thus, a single LM step requires the numerical solution of $m+1$ coupled systems, each one consisting of n ODEs. Fortunately, these systems are coupled in a special way, namely, for each $i = 1, 2, \ldots, m$, system (2.6) is a system of ODEs for $\frac{\partial \mathbf{y}}{\partial \theta_i}$, coupled only with (2.1). The system of equations (2.6) has the same stiffness as (2.1) and therefore the same step size can be used for the time integration of (2.1) and (2.6). Therefore, the one-way coupling can be used to solve (2.1) and (2.6) efficiently. Still, this approach has limitations for large scale problems due to computational costs.

Another approach to approximate the matrix $J(\theta)$ could be by means of divided differences instead of numerically solving (2.6). The j-th column of $J(\theta)$ is then given by

$$\frac{\partial \mathbf{Y}(\theta)}{\partial \theta_j} \approx \frac{\mathrm{Y}(\tilde{\theta}^j) - \mathrm{Y}(\theta)}{\delta\tilde{\theta}_j}, \tag{2.7}$$

where the vector $\tilde{\theta}^j$ is obtained by a small perturbation $\delta\tilde{\theta}_j$ in the j-th entry of θ. In this case, for one LM step system (2.1) has to be numerically integrated $m+1$ times. With regard to the computational costs, when $\mathbf{f}$ is nonlinear, it is more expensive than the previous approach where the linear systems of variational equations are solved. Moreover, the drawback of divided difference method is that the numerical approximations (2.7) introduce additional errors.

Remark 2.2 For large scale problems computation on a single computer can become unfeasible and one needs to use a parallel machine. Parallelization of the computational work when (2.1) and (2.6) are solved numerically is possible at the level of a time step of the time integrator. Therefore, it will be inefficient due to heavy communication. The advantage of the divided difference approach is that in this case (2.1) is solved for $m+1$ different values of θ independently of each other. Therefore, parallelization of the computational work is trivial and can be very efficient.

Remark 2.3 Given $\mathbf{f}$ and $\mathbf{y}_0$, the partial derivatives $\frac{\partial \mathbf{f}}{\partial \mathbf{y}}$, $\frac{\partial \mathbf{f}}{\partial \theta_i}$, $\frac{\partial \mathbf{y}_0}{\partial \theta_i}$ $(i = 1, \ldots, m)$ in (2.6) can be, in principle, found analytically. However, for large scale problems when $\mathbf{f}$ has a

complicated nonlinear form, this can be a tedious work to do. In such cases, these derivative functions can be generated automatically by using a symbolic mathematics package, like *Maple* [12] or *Mathematica* (Wolfram Research, Inc).

Remark 2.4 Numerical integration of (2.1) and (2.6) requires a fast and reliable ODE solver. Search in the parameter space may lead to some values of θ such that the systems of ODEs become stiff [9]. Therefore, an implicit scheme is the best choice for time integration both with respect to computational speed and for stability reasons. Moreover, using an implicit scheme allows us to exploit the specific coupling between (2.1) and (2.6) in an efficient way. At each time step integrating first (2.1) provides the solution vector $\mathbf{y}$ and the LU decomposition of the Jacobian matrix $I_m - \tau\frac{\partial \mathbf{f}}{\partial \mathbf{y}}$, where τ is the time step. Then the calculation of $\frac{\partial \mathbf{y}}{\partial \theta_i}$ from (2.6) reduces to a simple forward substitution and backsubstitution. In our simulations we use the implicit multistep Backward Differentiation Formulas (BDF) [13].

Remark 2.5 When the model includes algebraic equations, the systems of ODEs (2.1) and (2.6) change to Differential Algebraic Equations (DAEs). Since we use an implicit solver for the time integration, the method we have described here is readily applicable for that type of models.

Remark 2.6 When the unknown parameters have to obey certain constraints, linear or nonlinear, some additional work might be needed. If the correction $\delta\theta$ found by (2.5) leads to violation of some constraints, then by the introduction of Lagrange multipliers a modified correction can be found, which fits all constraints. For the constrained minimization problem we refer the reader to [14].

2.2 Statistical analysis of obtained parameters

Above we used θ^* to denote the 'true' parameter vector, for which (2.1) describes reality with sufficient accuracy, and by $\hat{\theta}$ we denote the parameter vector which minimizes (2.2). Remarkably, even having a 'right' model and an estimate $\hat{\theta}$ for the parameter vector which fits the data well, does not mean that the whole modelling problem is resolved successfully. It is important to know how reliable the obtained estimate is. In other words, we need information about the difference $\hat{\theta} - \theta^*$. In order to investigate the quality of the estimate $\hat{\theta}$, one needs to include some statistical analysis [9, 15, 16].

We assume that the measurement errors in $\tilde{y}_i$ are independent of each other and normally distributed and that the system (state vector $\mathbf{y}$) is well scaled, so that the error distributions have zero mean and constant standard deviation σ. Then, $\hat{\theta}$ is a maximum likelihood estimate [15]-[16]. By assumption the model with the 'true' solution θ^* describes reality, so

$$\tilde{y}_i \approx y_{c_i}(t_i, \theta^*) + \epsilon_i, \quad i = 1, 2, \ldots, N, \tag{2.8}$$

where ϵ_i are the measurement errors, for which

$$\hat{\theta} - \theta^* \sim N_m \quad 0, \sigma^2 \left(J^T(\hat{\theta}) J(\hat{\theta}) \right)^{-1} \tag{2.9}$$

holds approximately [15]. Here $N_m(\cdot, \cdot)$ denotes the m-dimensional multivariate normal distribution. Notice that (2.9) holds exactly when $\mathbf{y}$ is linear in θ. The $(1 - \alpha)$-confidence

region for θ^* is determined by the inequality

$$(\theta^* - \hat{\theta})^T \left(J^T(\hat{\theta}) J(\hat{\theta}) \right) (\theta^* - \hat{\theta}) \leq \frac{m}{N-m} S(\hat{\theta}) F_\alpha(m, N-m), \tag{2.10}$$

where $F_\alpha(m, N-m)$ is the upper α part of Fisher's distribution with m and $N-m$ degrees of freedom. For instance, with $\alpha = 0.05$ we have a 95% chance that θ^* lies in this region. This ellipsoidal confidence region allows us to assess the quality of the computed parameter vector $\hat{\theta}$. The ellipsoid defined by (2.10), is centered at $\hat{\theta}$ and has its principal axes directed along the eigenvectors of $J^T(\hat{\theta})J(\hat{\theta})$. Using the SVD (2.4) for $J(\hat{\theta})$, we get

$$J^T(\hat{\theta}) J(\hat{\theta}) = V(\hat{\theta}) \Sigma^2(\hat{\theta}) V^T(\hat{\theta}),$$

and the eigenvectors of $J^T(\hat{\theta})J(\hat{\theta})$ are the columns of the matrix $V(\hat{\theta})$. So, the ellipsoid has its principal axes directed along the column vectors of the matrix $V(\hat{\theta})$. Moreover, the radii along these principal axes are inversely proportional to the corresponding singular values σ_i, the diagonal elements of $\Sigma(\hat{\theta})$. This all can be seen by using the following transformation (rotation)

$$\mathbf{z} = V^T(\hat{\theta})(\theta^* - \hat{\theta}), \tag{2.11}$$

yielding

$$(\theta^* - \hat{\theta})^T \left(V(\hat{\theta}) \Sigma^2(\hat{\theta}) V^T(\hat{\theta}) \right) (\theta^* - \hat{\theta}) = \mathbf{z}^T \Sigma^2(\hat{\theta}) \mathbf{z} = \sum_{i=1}^{m} \sigma_i^2 z_i^2. \tag{2.12}$$

On the other hand, since $S(\hat{\theta})/(N-m)$ is an unbiased estimator of σ^2, the equation for the ellipsoid can be rewritten as

$$\sum_{i=1}^{m} \sigma_i^2 z_i^2 = r_\sigma^2, \tag{2.13}$$

where $r_\sigma^2 \approx m\sigma^2 F_\alpha(m, N-m)$ is proportional to the variance in the measurement errors. This form is more convenient to deal with because $\mathbf{z}$ can be considered as a set of uncorrelated variables, and once the conclusion has been drawn for the determinability of $\mathbf{z}$, the problem can be transformed back, revealing us the quality of $\hat{\theta}$.

Now, we assume that the model (2.1) is properly scaled, such that all parameter values are of the same order of magnitudes, and that we are interested only in the first few digits of the parameter values. Let us introduce the sphere given by

$$\sum_{i=1}^{m} z_i^2 = r_\epsilon^2, \tag{2.14}$$

where r_ϵ defines the level of accuracy one desires for the parameter estimates. For instance, if the parameters are of order $O(1)$ and one is interested only in the first two digits to the right of the decimal point, then $r_\epsilon = 0.01$. In order to be able to determine z_i accurately enough, the radius along the ellipsoid's i-th principal axis shouldn't exceed the radius of the sphere, which leads us to the following inequality

$$\sigma_i \geq \frac{r_\sigma}{r_\epsilon}. \tag{2.15}$$

A graphical representation of the ellipsoid and the sphere is given in Figure 2.1 for the 2-dimensional case. If only the first k largest singular values satisfy (2.15), then only the

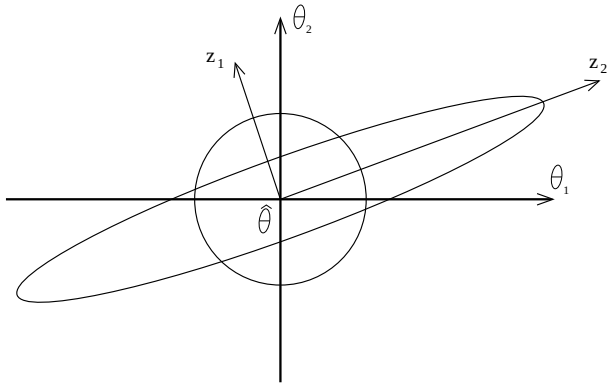

Figure 2.1: Example of an ellipsoidal confidence region and an accuracy sphere in the 2-dimensional case; clearly, z_1 is well-determined, while z_2 is not.

first k entries of $\mathbf{z}$ are estimated with the required accuracy and no sufficient information is available for the remaining components of $\mathbf{z}$. Now, recalling (2.11) and the fact that V describes a rotation around the center of the ellipsoid, it becomes clear that only the set of the first k largest singular values contain useful information about the quality of the parameter estimates. Each corresponding eigenvector defines a parameter or a linear combination of parameters which is well-determined. In the case when a principal axis of the ellipsoid makes a significant angle with the axis in parameter space (i.e., there exists more than one significant entry in the eigenvector), this corresponds to the presence of correlation among parameters in $\hat{\theta}$. The remaining degrees of freedom in the parameters, corresponding with the smaller singular values, cannot be determined (with sufficient accuracy) by means of the available experimental data.

To summarize, the level of noise in the data in combination with the accuracy requirement for the parameter estimates, defines the threshold for significant singular values in the matrix Σ. The number of singular values exceeding this threshold determines the number of parameter relations that can be derived from the experiment. How these relations relate to the individual parameters is described by the corresponding columns in the matrix V. The largest entries in these columns indicate the well-determined parameters and, on the other hand, if entries are small, then the corresponding parameters cannot be determined with reasonable accuracy.

From (2.10) one can also derive dependent confidence intervals for the parameter estimates, which are the intersections of the ellipsoidal region with the parameter axes

$$\theta_i : |\theta_i - \hat{\theta}_i| \leq r_\sigma \quad \sqrt{\left(V(\hat{\theta})\Sigma^2(\hat{\theta})V^T(\hat{\theta})\right)^{-1}_{ii}} \quad , \quad i = 1, 2, \ldots, m, \tag{2.16}$$

and independent confidence intervals, which are the projections of the ellipsoidal region on the parameters axes

$$\left\{\theta_i : |\theta_i - \hat{\theta}_i| \leq r_\sigma \quad \sqrt{\left(V(\hat{\theta})\Sigma^{-2}(\hat{\theta})V^T(\hat{\theta})\right)_{ii}}\right\}, \quad i = 1, 2, \ldots, m. \tag{2.17}$$

Clearly, small independent confidence intervals for $\hat{\theta}_i$ indicate that it is well-determined. However, in some cases considering only individual confidence intervals can be misleading. For instance, in the presence of a strong correlation between parameters, the dependent confidence intervals underestimate the confidence region while the independent confidence intervals overestimate it.

Finally, (2.13) indicates that having, for instance, two times more accurate data so that the standard deviation σ is halved, will decrease the radii along the ellipsoid's principal axis by a factor of 2. Therefore, in case of very small singular values σ_i (i.e. strongly elongated ellipsoids) more accurate data obtained by the experimentalist will not improve much the quality of the corresponding parameter estimates. In such a case, one certainly needs additional measurements of a different type (e.g., different components, different time points, or in the case of PDEs different spatial points).

3 A large-scale biological test problem

In this section we study the model of the genetic regulatory network at the early stage of development of *Drosophila melanogaster*. In particular, we are interested in the spatio-temporal pattern formation of gap gene expression in the *Drosophila* embryo during the early cleavage cycles 13 and 14A. The gap gene system includes the genes Bicoid (*bcd*), Caudal (*cad*), Hunchback (*hb*), Kruppel (*Kr*), Knirps (*kni*), Giant (*gt*) and Tailless (*tll*). It is known that before cycle 13 there is no (significant) expression of gap genes in the embryo. The process of pattern formation for gap gene expression is initiated by gradients of the maternal proteins *bcd*, *hb* and *cad*. The size, location and dynamics of gap domains depend on regulatory interactions between the genes involved in the system. This regulatory network is well studied in [5]-[6]. There, a global search approach based on simulated annealing (SA) is used for the estimation of the parameters in the gap gene model. A more efficient approach, namely combining a global search method, the Stochastic Ranking Evolution Strategy (SRES), with a local direct search method, Downhill Simplex (DS), is introduced in [7]. The quality of the parameter estimates is measured by the root mean square (RMS) of the discrepancy vector and considered to be 'good' if $RMS < 12.0$ and if there are no specific pattern defects in the model response [5]-[7]. As explained in the previous section we should notice that this definition of the quality of parameter estimates can be rather misleading. In fact, RMS shows the quality of the fit of the model response to the data but does not give any information about the quality of the parameter estimates. Our aim is to find the parameter estimates that give a good fit and to apply statistical analysis in order to investigate how reliable these estimates are.

3.1 The mathematical model

We first outline the main aspects of the mathematical model which is used to describe the mechanism of pattern formation at the early developmental stage of the *Drosophila* embryo. Detailed information can be found in [4]-[6]. The change of the level of concentrations of gene products is described by the system of ODEs

$$\frac{dg_i^a}{dt} = R_a \Phi \left(\sum_{b=1}^{N_g} W_a^b g_i^b + m_a g_i^{bcd} + h_a \right) - \lambda_a g_i^a + D_a \left(g_{i+1}^a - 2g_i^a + g_{i-1}^a \right), \qquad (3.1)$$

where a and b denote gene products, g_i^a denotes the concentration of gene product a at nucleus i, g_i^{bcd} denotes the concentration of maternal protein *bcd* (constant in time) at nucleus i, $N_g = 6$ is the number of genes, and the function

$$\Phi(x) = \frac{1}{2}\left(\frac{x}{\sqrt{x^2+1}} + 1\right) \tag{3.2}$$

is a sigmoid function. Note that indexes a and b used in (3.1) are integers. To avoid misunderstanding genes *cad*, *hb*, *Kr*, *kni*, *gt*, *tll* are enumerated from one to N_g, respectively. Indexes with integers and abbreviations of genes are used here interchangeably. For instance, D_2 is the same as D_{hb}.

In the system (3.1) there are in total $m = 66$ unknown parameters. These include the regulatory weight matrix W of size $N_g \times N_g$ with the entries W_a^b representing the regulation of gene a by gene b, maternal coefficients m_a representing the regulatory effect of *bcd* on gene a, promoter thresholds h_a, promoter strengths R_a, diffusion coefficients D_a, and decay rates λ_a.

Since the nuclei are equally distributed along the anterior-posterior (AP) axis of the embryo, (3.1) can be seen as a discretized (in space) form of a system of one-dimensional reaction-diffusion equations. The region of interest includes 30 and 58 nuclei at the central part of the embryo during the cycles 13 and 14A, respectively. Therefore, there are 180 and 348 equations in the system (3.1) at the cycles 13 and 14A, respectively. Initial conditions at $t = 0.0$ (beginning of cycle 13) are prescribed by gradients of *hb* and *cad* and zero levels for the other genes. The model simulates until gastrulation at $t = 71.1$. At the boundaries the central difference in the last term in the right-hand side of (3.1) is replaced by a one-sided difference (no-flux conditions).

During the mitosis phase between cycles 13 and 14A (see Figure 3.1) the protein production in the embryo is shut down and therefore the first term in the right hand side of (3.1) does not contribute anything. Mitosis starts at $t = 16.0$ and ends at $t = 21.1$. At the end of the mitosis all nuclei simultaneously divide. This is done by doubling the number of nuclei, dividing diffusion coefficients by 4 so that the distance between nuclei is halved, and copying the concentration values from each nucleus to its daughter nuclei. The latter provides the initial conditions for equations (3.1) in cycle 14A.

3.2 The data

The data set, consisting of $N = 2702$ measurements, is available from the FlyEx database [8]. The level of measurement error is less than 5%, see [17]. Figure 3.1 shows the time points T_i $(0 \le i \le 8)$ when measurements were taken. Figure 3.2 shows the gene expression data

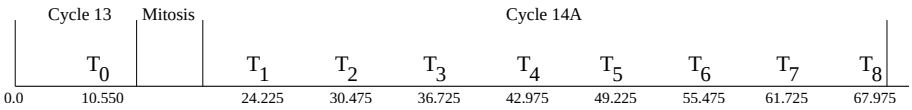

Figure 3.1: Time axis and the points when measurements were taken: one in cycle 13 and eight in cycle 14A; mitosis is the phase between two cycles when there is no protein production in the embryo.

at time points T_i $(0 \leq i \leq 8)$. Note that measurements for the concentrations of all gene products at all time points are available, except *cad* at T_7, T_8 and *tll* at T_0, T_1, T_2.

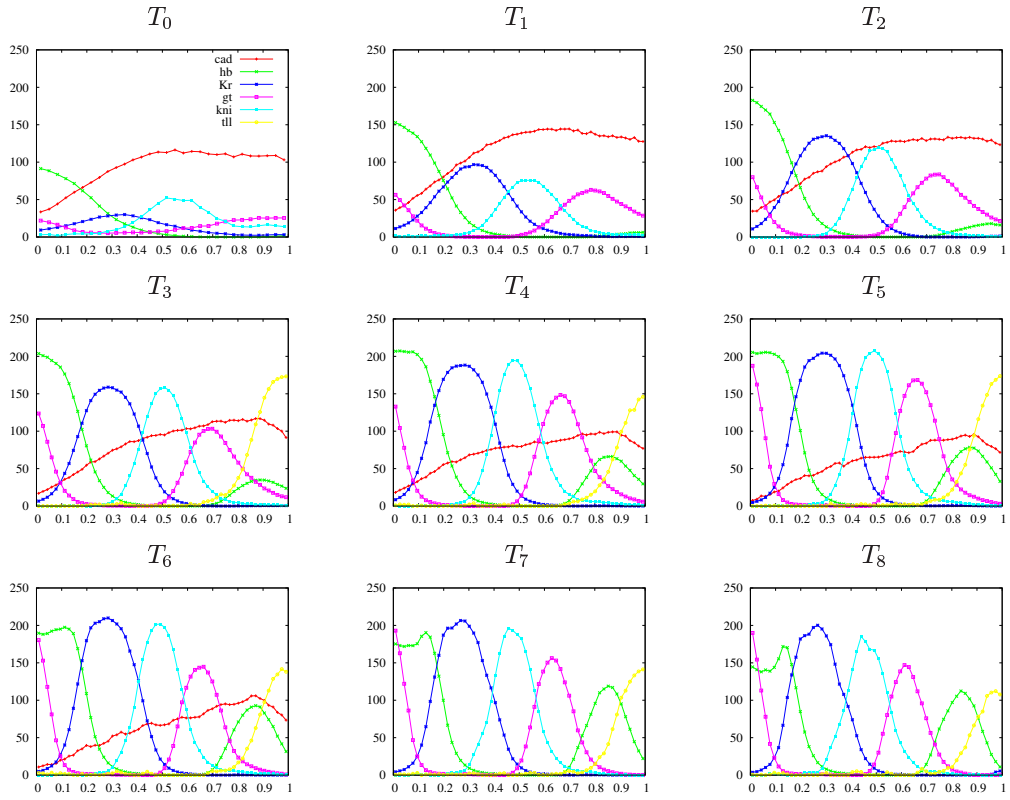

Figure 3.2: Gene expression data at different time points. Graphs show relative protein concentration (with a range from 0 to 255 fluorescence units) plotted against position on the AP axis (the region of interest is scaled to $[0, 1]$).

It is not known how the experimental errors are distributed. However, measurement values are the outcomes of sophisticated data processing procedures, see [17]. In fact, all data points are values integrated over space and averaged over the number of individual embryos (the number of embryos varies from 9 to 62 for different time points and different genes [8]). Therefore, from a statistical point of view, it is reasonable to assume that the experimental errors are normally distributed.

3.3 The experimental setup

We apply the LM method to estimate the parameters for the gap gene system. Since the LM method is a local search approach, the choice of initial values for the parameter vector is important for convergence. Fortunately, for this problem there is extensive information available in the literature. We use 80 different initial values for the parameter vector θ

from [7]. Each of these parameter sets is obtained by using an evolution strategy (global approach) combined with direct search (local approach).

With the notations introduced in Sections 3.1-3.2, RMS is defined as

$$RMS(\theta) = \sqrt{\frac{1}{N}\sum_{a=1}^{N_g}\sum_{i=1}^{N_c}\sum_{j=0}^{8}\alpha_j^a\left(g_i^a(T_j,\theta)_{model} - g_i^a(T_j)_{data}\right)^2},$$

where N_c is the number of nuclei and α_j^a is equal to zero for Tll at $j = 0, 1, 2$ and for cad at $j = 7, 8$, and is equal to one otherwise. We note that only 41 of the initial parameter sets have $RMS(\theta) < 12.0$, see Table 3.1.

The search space for parameters is defined by the linear constraints

$$10.0 \le R_a \le 30.0, \quad 0.0 < D_a \le 0.3, \quad 5.0 \le \frac{ln(2)}{\lambda_a} \le 20.0, \quad a = 1, \ldots, N_g, \tag{3.3}$$

and by the nonlinear constraints

$$\sum_{b=1}^{N_g}\left(W_a^b g_{max}^b\right)^2 + \left(m_a g_{max}^{bcd}\right)^2 + (h_a)^2 \le 10^4, \quad a = 1, \ldots, N_g, \tag{3.4}$$

where g_{max}^b and g_{max}^{bcd} are the maximum values in the data set for gene b and protein bcd, respectively. Note that in [5]-[7] threshold parameters h_a for genes Kr, Kni, gt, and hb are fixed to negative values representing a constitutively repressed state for the corresponding genes [18]. Fixing some parameters to specific values may severely restrict the search space leaving some solutions out of consideration. Contrary to their approach, we include threshold parameters for these genes in the search by putting the constraints $-10.0 \le h_a \le 0.0$.

In order to make the analysis of parameter estimation easier, we scale in advance all parameters used in (3.1) in the following way:

$$\tilde{R}_a = 0.1R_a, \quad \tilde{D}_a = 10D_a, \quad \tilde{\lambda}_a = 10\lambda_a, \quad \tilde{W}_a^b = 10^2 W_a^b, \quad \tilde{m}_a = 10^2 m_a, \quad \tilde{h}_a = h_a,$$

for all genes a and b. Note that the choice of the scaling factors for R_a, D_a, and λ_a is based on the search ranges of the corresponding parameters. The choice of the scaling factors for regulatory weights W_a^b and maternal coefficients m_a is based on the fact that the maximum level of protein concentration for all genes in the data set is of order $O(10^2)$. Thus, all scaled parameters are of order $O(1)$.

3.4 Results of parameter estimation

The least squares estimation using the LM method yields a significant decrease of RMS in all simulations, see Table 3.1. There are only 5 initial parameter sets having $RMS < 10.0$, with the best fit having $RMS = 9.56$. After using the LM method there are 71 final parameter sets which have $RMS < 10.0$ and among them there are 69 having values of RMS uniformly distributed between 8.37 and 9.43. It is difficult to make a distinction between these 69 parameter estimates based only on RMS values. Therefore, in our analyses, we take into account all of them. We note that there is a distinct gap between the RMS values of the chosen 69 parameter sets and the RMS values of the remaining parameter estimates.

	$RMS < 10.0$	$10.0 \leq RMS < 12.0$	$12.0 \leq RMS < 14.0$	$RMS \geq 14.0$
θ^{in}	5	36	21	18
$\hat{\theta}$	71	3	1	5

Table 3.1: Numbers in the table show the number of parameter estimates with corresponding ranges for RMS, where θ^{in} and $\hat{\theta}$ correspond to the parameter estimates before and after using the LM method, respectively.

Parameter estimates found by the LM method also produce a better fit than those previously obtained in [5]-[7]. In Figure 3.4 the model response for one of our parameter sets (green lines) is compared to the data (red lines) and to the patterns obtained with the parameter set from [5] (blue lines). The patterning defects reported in [5], such as the expression of hb at the anterior and posterior borders, are mainly resolved. However, there are two problems, mentioned in [5]-[6], that remain unsolved with the new parameter estimates. The first one is related to the artificially high level of gap gene expression at cycle 13, i.e the model responses are much larger than the data values yielding large positive discrepancies. This is apparently due to the model itself. It might be needed to include some delay in the process of protein production to be able to overcome the poor fit at cycle 13, as it is proposed in [5]-[6]. The second one is related to the absence of boundary shifts for the posterior hb domain in the model responses.

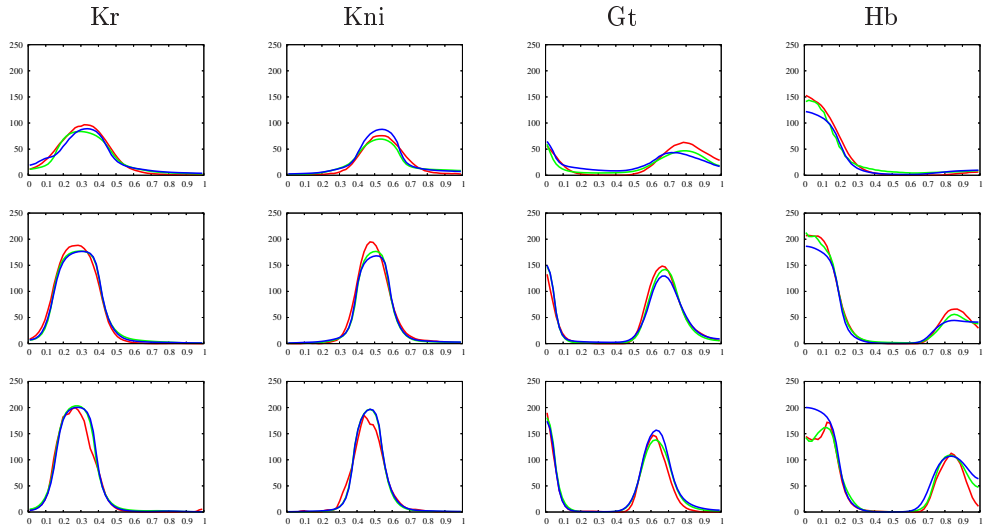

Figure 3.3: Comparison between data (red lines), patterns obtained by parameter set from [5] (blue lines) and patterns with the parameter set yielded from the LM search (green lines) for the expression of gap genes Kr, Kni, gt, and hb at early ($t = 24.225$, first row) mid- ($t = 42.975$, second row) and late ($t = 67.975$, last row) cycle 14A. Axes are as in Figure 3.2.

Information about the regulatory matrix for all parameter sets is given in Table 3.2. Triplets show the number of parameter sets in which a regulatory weight falls into one of the

following categories: repression (values ≤ -0.005)/ no interaction (values between -0.005 and 0.005)/ activation (values ≥ 0.005). Based on the highest value in the triplets, the table is coloured such that the background colours represent activation (green), no interaction (light-blue), or repression (pink).

	bcd	*cad*	*hb*	*Kr*	*gt*	*kni*	*tll*
cad	60/5/4	69/0/0	69/0/0	69/0/0	69/0/0	69/0/0	69/0/0
hb	0/0/69	0/1/68	0/1/68	3/57/9	3/29/37	69/0/0	4/41/24
Kr	0/0/69	0/0/69	24/45/0	0/4/65	67/1/1	43/26/0	69/0/0
gt	0/0/69	0/0/69	7/47/15	69/0/0	0/0/69	0/11/58	45/24/0
kni	5/4/60	0/7/62	69/0/0	38/31/0	51/18/0	0/0/69	67/2/0
tll	42/7/20	12/6/51	42/16/11	64/3/2	60/4/5	67/2/0	0/11/58

Table 3.2: Maternal coefficients and regulatory weight matrix for the gap gene system based on 69 parameter sets found by the LM method. Numbers show how many parameter sets have repression / no interaction / activation for corresponding regulatory weight. Colours indicate activation (green), no interaction (light-blue), or repression (pink) based on the maximum values in triplets.

Our results are in good agreement with the results obtained in [5]-[7]. Namely,

- *cad* and *bcd* activate gap genes hb, Kr, gt, and kni;
- gap genes hb, Kr, gt, and kni have autoactivation;
- terminal gap gene tll represses gap genes Kr, gt, and kni;
- mutually exclusive gap genes strongly repress each other, these correspond to weights W_{gt}^{Kr}, W_{Kr}^{gt}, W_{hb}^{kni}, and W_{kni}^{hb};

Previous results also suggest that pairs of overlapping gap genes, namely, hb and gt, hb and Kr, gt and kni, Kr and kni, either have no interaction with each other or repress each other, except for the effect of gt on hb, see [5]. These regulations are partially confirmed here. However, we find that the effect of Kr on hb and hb on gt can be positive as well in some cases. A striking difference is that kni mostly activates gt while previously it was found that there was no interaction between them.

Scatter plots in Figure 3.4-3.5 show the range of the parameter estimates for the gap gene system. For each individual parameter indicated on the horizontal axis, its estimated values (red circles) are plotted along the vertical axis. Most of the parameters have a broad range of possible values, meaning that they are not uniquely found. The only exceptions are some entries in the regulatory weight matrix, such as W_{gt}^{hb}, W_{hb}^{Kr}, W_{gt}^{kni}, and W_{hb}^{tll}.

3.5 Determinability of parameters

We apply the statistical analysis introduced in Section 2 to all 69 parameter sets obtained by the LM method to asses the quality of estimates.

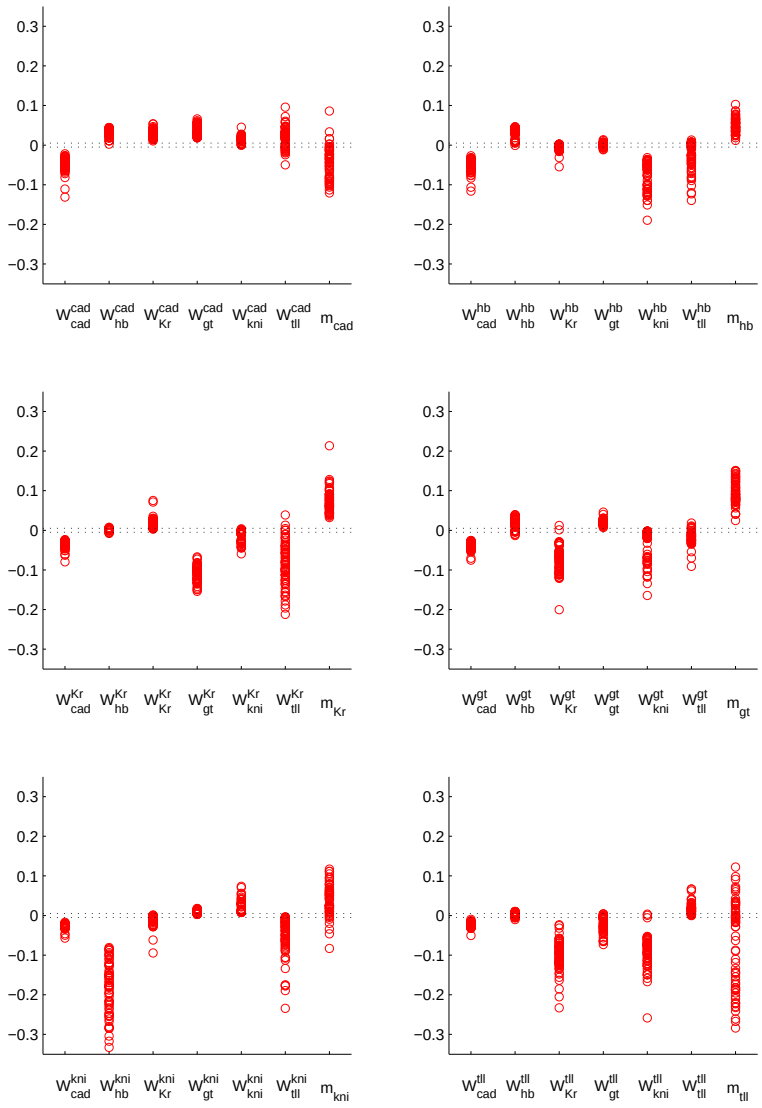

Figure 3.4: Scatter plots of parameters in the regulatory weight matrix for the gap gene system.

Ellipsoidal confidence regions corresponding to parameter estimates are given by (2.10). A trivial check reveals that none of the parameter estimates lies in the ellipsoidal confidence regions of all other parameter sets. Note that this does not necessarily imply that there are 69 different minima or solutions for the parameter vector.

Dependent and independent confidence intervals for each parameter set can be computed by (2.16) and (2.17), respectively. We check if the corresponding confidence intervals fall into the repression, no interaction, or activation category. Colours in Table 3.2 do not

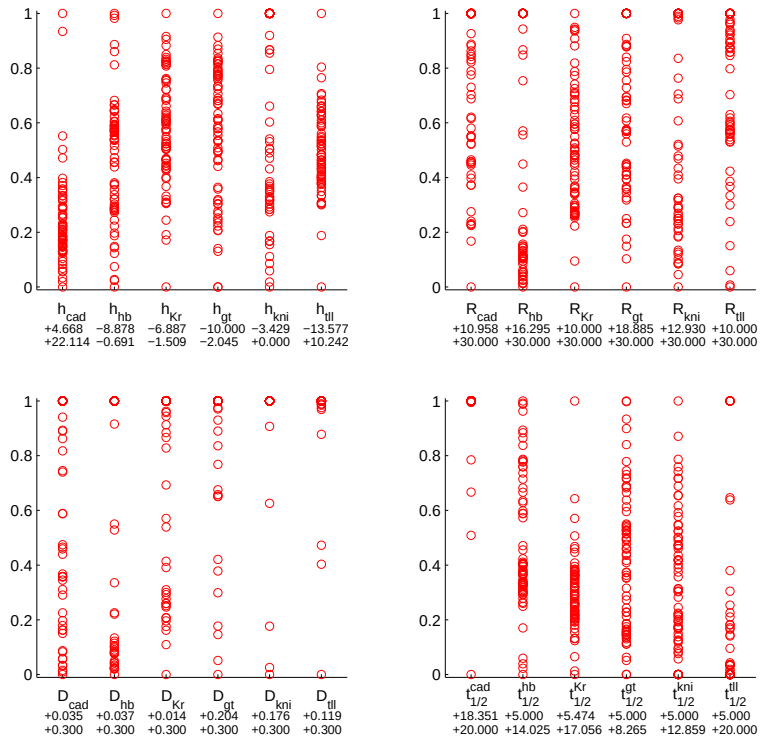

Figure 3.5: Scatter plots of parameters h, R, D and $t_{1/2} = ln(2)/\lambda$.

change when only dependent confidence intervals are taken into account. However, including independent confidence intervals one can no longer make any qualitative conclusions about the entries in the regulatory weight matrix. So, individual confidence intervals are not informative for our purpose.

For each parameter set $\hat{\theta}$, the SVD (2.4) of the Jacobian $J(\hat{\theta})$ yields the matrices $V(\hat{\theta})$ and $\Sigma(\hat{\theta})$. In order to find the number of singular values in $\Sigma(\hat{\theta})$ satisfying (2.15) we need to quantify r_σ and r_ϵ. We are interested only in the first digit to the right of the decimal point of the scaled parameters and therefore we take $r_\epsilon = 0.1$. Since $\sigma \approx \frac{S(\hat{\theta})}{N-m} = \sqrt{\frac{N}{N-m}}\, RMS(\hat{\theta})$, we have

$$r_\sigma \approx \sigma\sqrt{m\ F_\alpha(m, N-m)} \approx \sqrt{\frac{N\ m}{N-m}\ F_\alpha(m, N-m)}\ RMS(\hat{\theta}).$$

For $\alpha = 0.05$ we then obtain $r_\sigma \approx 9.4\ RMS(\hat{\theta})$ (the choice of α does not make much difference here due to the large value of N).

Investigation of all parameter sets shows that, on average, 15 singular values satisfy (2.15) meaning that at most 15 parameters or linear combinations of them can be determined with two digits accuracy. There is a set of parameters which have significant entries in the first 15 columns of all V matrices. It includes regulatory weights W^{cad}_{Kr}, W^{cad}_{gt}, W^{cad}_{kni}, W^{cad}_{tll}, W^{hb}_{Kr}, promoter thresholds h_{Kr}, h_{gt}, h_{tll}, decay rate λ_{cad}, and promoter strength R_{kni}. However,

inspection of the first 15 columns of the V matrices shows that there is not a single parameter which can be determined individually. It means that a principal axis of the ellipsoid makes an angle with the corresponding axes in parameter space. The same holds for other principal axes of the ellipsoid defined by the columns of the matrix V corresponding to singular values which do not satisfy (2.15).

Let us investigate here possible reasons for correlations among parameter estimates. Lack of accuracy of the data cannot be the reason for that. More accurate data would simply make the ellipsoid shrink but not rotate and therefore it would not improve the determinability of parameters.

We have also checked whether data insufficiency may cause the nondeterminability of parameters. This is done in the following way. Assume that a larger data set was available, say we had measurements for all gene products, in all nuclei, at 71 uniformly distributed time points. With these choices the total number of measurements would be $N = 21180$. Since the Jacobian depends only on the model responses and not on the values of the data, we can generate a new Jacobian $\tilde{J}(\hat{\theta})$ including all 'ghost' data points. From the SVD of the corresponding $\tilde{J}(\hat{\theta})$ we get the matrices $\tilde{V}(\hat{\theta})$ and $\tilde{\Sigma}(\hat{\theta})$ which define new ellipsoidal regions. The ellipsoids are slightly rotated in comparison with the initial ones but not enough to make the principal axes of the ellipsoid get closer to the parameter axes. From this we conclude that lack of data points is not the reason for these correlations.

Correct application of the statistical analysis for the parameter estimates described in Section 2 implies that the measurement errors are independent and come from a normal distribution. To study whether these assumptions affect the determinability we conduct the inverse experiment. We take one of the parameter sets obtained by the LM search, having $RMS = 8.38$, and we denote it by θ^*. By integrating the model equations with θ^* we generate an exact data set at the same data points as the initial data set. To the exact data values we add errors drawn from the normal distribution with zero mean and standard deviation equal to 8.5. From the exact and the perturbed data set, we compute $RMS(\theta^*) = 8.17$. The perturbed data set is used for the parameter estimation by means of the LM search. Note that by constructing this inverse problem, we make sure that the assumptions about the measurement errors are correct. With 40 different initial values of θ from [7] we obtain 34 parameter estimates having RMS between 7.95 and 8.25. The ranges of the values of the obtained parameters remain broad (data is not shown here). Inspection of the corresponding V matrices shows that parameters are not determinable due to the correlations, similar to the original problem.

We conclude that the observed correlations among parameters are a property of the model. Since an explicit form of the dependence of the state vector on the parameters is not known, the use of reparametrization techniques is not feasible. Note that the majority of parameters in (3.1) appear in the argument of sigmoid function Φ. If the model (3.1) is used to obtain only the qualitative information, such as the signs of regulatory weights, then the particular mathematical form of this function is of no importance [4]. However, it has to be studied if the choice of the sigmoid function affects the determinability of parameters. Preliminary results suggest that the correlations among parameters are reduced when the sigmoid function defined by (3.2) is replaced by a piecewise linear function.

To summarize, the statistical analyses show that parameters in (3.1) cannot be determined individually due to the correlations among the parameters. The observed correlations are a property of the model itself, not the data. Further investigation is needed to study the model equations to remove the correlations so that the parameters can be well determined.

Finally, we remark that the statistical analysis, introduced in Section 2, has been derived for models that are linear in θ. In the nonlinear case, it holds approximately and the accuracy of the approximation depends on the type of nonlinearity. Obviously, the solution of (3.1) is nonlinear in θ and therefore all conclusions which are drawn here are approximate in that sense.

4 Concluding remarks

In this paper we have applied the Levenberg-Marquardt (LM) optimization method to estimate the parameters in the model of the genetic regulatory network in *Drosophila* embryo. Statistical analysis is used to study the quality of the obtained parameter estimates, i.e. how well the parameters are determined with the available experimental data.

The parameter estimates obtained with the LM method fit the data better than the parameters known from literature. For instance, the defects in the patterns of gene product of *Hunchback*, reported in [5], are removed here. Qualitative conclusions with the new parameter sets are in good agreement with the results stated previously. Namely, the regulatory interactions among genes involved in gap gene system are confirmed here, with only one exception. We found that gene *Knirps* activates gene *Giant* while previously it was stated that there was no interaction between them. Large ranges for the values of parameter estimates suggest that the parameters are not unique.

Determinability studies based on statistical analysis show that the model cannot be used as a quantitative tool. None of the parameters used in the model can be determined individually due to correlations among parameters. We have shown that these correlations are not related to a lack of data. The nondeterminability stems from the intrinsic correlations among parameters in the model. Further investigation is needed to modify the model equations in order to remove the observed correlations.

The products of mathernal genes regulate gap genes, but not vice versa [19]. Determinability studies based on statistical analysis suggest that the model (3.1) can be reduced in size with respect to the number of equations and the number of parameters, by removing the model equations for gene *cad*. However, this is an open question for further work.

Acknowledgment We acknowledge support from the Dutch BSIK/BRICKS project and from NWO's 'Computational Life Science' program, projectnr. 635.100.010. We would like to thank prof. P. W. Hemker and prof. J. G. Verwer for their valuable comments and suggestions.

References

[1] J. D. Murray (2002), *Mathematical Biology*, Springer, Berlin.

[2] J. Nocedal, S. J. Wright (1999), *Numerical Optimization*, Springer, New York.

[3] S. B. Carroll, J. K. Grenier, S. D. Weatherbee (2001), *From DNA to Diversity*, Blackwell Science, Malden, Massachusetts.

[4] J. Reinitz, D. H. Sharp (1995), *Mechanism of eve stripe formation*, Mech. Dev. 49 (1-2), pp. 133-158.

[5] J. Jaeger, J. Reinitz (2004), *Dynamical analyses of regulatory interactions in the gap gene system of Drosophila melanogaster*, Genetics 167, pp. 1721-1737.

[6] J. Jaeger, J. Reinitz (2004), *Dynamic control of positional information in the early Drosophila embryo*, Nature 430, pp. 368-371.

[7] Yves Fomekong Nanfack, Jaap A. Kaandorp, Joke Blom (2007), *Efficient parameter estimation for spatio-temporal models of pattern formation: Case study of Drosophila melanogaster*, Bioinformatics 23, pp. 3356-3363.

[8] E. Poustelnikova, A. Pisarev, M. Blagov, M. Samsonova, and J. Reinitz (2004). *A database for management of gene expression data in situ*, Bioinformatics 20, pp. 2212-2221. (http://flyex.ams.sunysb.edu/flyex)

[9] P. W. Hemker (1972), *Numerical methods for differential equations in system simulation and in parameter estimation.*, Analysis and Simulation of Biochemical Systems, pp. 59-80.

[10] D. W. Marquardt (1963), *An algorithm for least-squares estimation of nonlinear parameters*, SIAM J. Appl. Math. 11, pp. 431-441.

[11] J. C. P. Bus, B. van Domselaar, J. Kok (1975), *Nonlinear least squares estimation*, CWI report, NW 17/75.
(http://repos.project.cwi.nl:8888/cwi_repository/docs/I/09/9052A.pdf)

[12] B. W. Char, K. O. Gedded, G. H. Gonnet, B. L. Leong, M. B. Monagan, S. M. Watt (1991), *Maple V Library Reference manual*, Springer Verlag.

[13] C. W. Gear (1971), *Numerical initial value problems in ordinary differential equations*, Prentice Hall, Englewood Cliffs.

[14] W. Stortelder (1998), *Parameter Estimation in Nonlinear Dynamical Systems*, PHD thesis, University of Amsterdam. (http://www.cwi.nl/~gollum/Stortelder_thesis.pdf)

[15] G. A. F. Seber, C. J. Wild (1988), *Nonlinear regression*, John Wiley & Sons, Inc.

[16] N. R. Draper, H. Smith (1981), *Applied regression analysis*, John Wiley & Sons, Inc.

[17] E. Myasnikova, A. Samsonov, M. Samsonov, J. Reinitz (2001), *Registration of the expression patterns of Drosophila segmentation genes by two independent methods*, Bioinformatics 17, pp. 3-12.

[18] J. Jaeger (2007), private communication.

[19] J. Jaeger & J. Reinitz (2006), *On the dynamic nature of positional information*, BioEssays 28, pp. 1102-1111.